Diversité lexicale et longueur du texte en évaluation du langage

Yves Bestgen

Université catholique de Louvain – yves.bestgen@uclouvain.be

Abstract

Estimating the lexical diversity of a text has been one of the main subjects of study in lexicometrics since its beginnings. It is also a question that has seen a large number of applications, particularly in the field of language acquisition and deterioration. In this field, indices based on probability laws, such as Muller's index, are considered too affected by length differences to be recommended, even though these indices are insensitive by construction. The aim of this paper is to understand why this research has led to an erroneous result, and to present a study carried out on 600 texts, written by learners of German, Czech and Italian, in order to rigorously evaluate the lexical diversity indices used in this field.

Keywords: Lexical diversity, random sampling, intra-class correlation coefficient, individual profiles.

Résumé

Estimer la diversité lexicale d'un texte est un des sujets principaux d'étude de la lexicométrie depuis sa naissance. C'est aussi une question qui a connu un grand nombre d'applications, tout particulièrement dans le domaine de l'acquisition et de la détérioration du langage. Dans ce domaine, des indices basés sur des lois de probabilité, comme l'indice de Muller, sont considérés comme trop affectés par les différences de longueur pour être recommandés. Il est pourtant bien établi que ces indices sont insensibles par construction. Cette communication a pour objectif de comprendre pourquoi ces recherches ont abouti à un résultat erroné et de présenter une étude menées sur 600 textes, rédigés par des apprenants de l'allemand, de l'italien et du tchèque, afin d'évaluer d'une manière rigoureuse les indices de diversité lexicale employés dans ce domaine.

Mots clés : Diversité lexicale, échantillonnage aléatoire, coefficient de corrélation intra-classe, profils individuels.

1. Introduction

Estimer la diversité lexicale (DL) d'un texte, c'est-à-dire le nombre de mots différents qu'il contient, est un des principaux champs d'étude de la lexicométrie depuis sa naissance. La raison majeure de cet intérêt, en plus du fait qu'il s'agit d'une question évidemment pertinente, est qu'y répondre est bien plus complexe qu'on pourrait le penser a priori. Un des premiers résultats de la lexicométrie a en effet été de montrer que le nombre de mots différents présents dans un texte était affecté d'une manière non linéaire par sa longueur, c'est-à-dire par le nombre total de mots contenus. De très nombreuses recherches ont été menées afin d'identifier des indices de DL qui permettent la comparaison de textes de longueurs différentes (voir Cosette (1994) et Tweedie et Baayen (1998) pour des synthèses). Ces travaux sont d'autant plus importants que les indices de DL sont employés dans nombre de recherches en linguistique appliquée, tout particulièrement dans le domaine de l'apprentissage et de la détérioration du langage comme en atteste les ouvrages de synthèses publiés sur cette question (Jarvis et Daller, 2013; Malvern et al., 2004).

Le plus étonnant lorsqu'on analyse cette littérature en évaluation du langage est de constater combien d'études ont été consacrées à proposer de nouveaux indices, à les valider dans diverses

situations d''évaluation et à essayer de déterminer quelle plage de longueur de textes (p.ex. de 150 à 350 mots) peut être analysée avec quel indice de sorte que les différences de longueur n'affectent pas les conclusions (p. ex., Fergadiotis, Wright et Green, 2015; Hess, Haug et Landry, 1989; Koizumi et In'nami, 2012; Malvern et al., 2004; McCarthy et Jarvis, 2007; Nasseri et Thompson, 2021; Treffers-Daller et al., 2018; Zenker et Kyle, 2021). Ces travaux ont conduit au développement d'indices comme MSTTR (Mean Segmental TTR), MATTR (Moving-Average Type-Token Ratio) et MTDL (Measure of Textual Lexical Diversity).

L'étonnement vient de ce qu'un indice insensible à la longueur a été proposé en lexicométrie dès 1964 par Charles Muller (1964a, 1964b). Cet indice utilise la loi binomiale afin de déterminer le nombre de mots différents que contiendrait chaque texte à comparer si sa taille était réduite à une longueur donnée, inférieure au plus petit de ces textes. Comme l'ont confirmé Hubert et Labbé (1988), Cossette (1994) et Baayen (2001), cet indice est une excellente approximation de l'indice que l'on peut dériver de la loi hypergéométrique, loi plus adaptée au problème que la loi binomiale. Cet indice peut être vu comme une généralisation de deux indices classiques, le K de Yule et le D de Simpson. En cela, il répond aux critiques de Thoiron (1986) à propos de la sensibilité excessive de D aux mots les plus fréquents. En linguistique appliquée, la version hypergéométrique de cet indice a été redécouverte par McCarthy et Jarvis (2007; voir aussi Serant, 1988) lors de leur étude approfondie de *vocd*, l'indice proposé par Malvern et al. (2004).

L'étonnement est encore plus grand lorsqu'on constate qu'une série de recherches en linguistique appliquée ont conclu que HD-D, et donc l'indice de Muller, était trop affecté par les différences de longueur pour être recommandé (Fergadiotis, Wright et Green, 2015; Koizumi et In'nami, 2012; Treffers-Daller, 2013; Zenker et Kyle, 2021). Comprendre pourquoi ces études ont abouti à ce résultat erroné et évaluer adéquatement les indices de DL employés dans ce domaine de recherche sont les deux objectifs de cette communication. La présente étude s'appuie sur une revue méthodologique sous presse (Bestgen, 2024). Si la méthodologie employée est similaire, les corpus analysés sont différents de même que la moitié des indices. La présente communication met aussi l'accent sur les travaux séminaux en lexicométrie.

2. Comment évaluer la sensibilité d'un indice de DL à la longueur des textes?

L'approche la plus évidente pour évaluer la sensibilité d'un indice de DL à la longueur d'un texte consiste à comparer des extraits de longueurs différentes, mais dont le contenu lexical est le plus proche possible. Presque toutes les recherches qui ont suivi cette voie ont employé la méthode d'échantillonnage parallèle popularisée par Hess, Haug et Landry (1989). Elle consiste à ramener tous les textes à une même longueur de départ en éliminant les mots situés au-delà d'une longueur arbitraire. La DL pour ce texte tronqué est comparée à la DL moyenne des segments d'une longueur donnée, obtenue en divisant le texte en 2, 3, ou 4. Cette méthode permet de comparer des segments de longueurs différentes, mais qui, lorsqu'on additionne tous les segments d'une même taille, contiennent exactement les mêmes mots. Si toutes les valeurs d'un indice sont très proches, cet indice est considéré comme non affecté par la longueur des segments. Pour tester cette condition, une analyse de la variance pour mesures répétées (les différentes longueurs) est employée et un effet non significatif de la longueur est recherché (Xanthos, 2013).

Le problème majeur posé par cette méthode est qu'elle donne l'illusion que les segments de différentes tailles que l'on compare ont le même contenu lexical. C'est n'est pas le cas parce que la répétition d'un mot à longue distance est prise en compte par les indices lorsqu'on analyse le

texte complet alors que c'est impossible dans les segments plus petits. La seule solution acceptable est d'employer une approche qui garantit que tous les mots aient la même probabilité d'être présents avec n'importe quel autre mot dans un échantillon d'une taille donnée. C'est évidemment le cas lorsque les extraits comparés sont obtenus par un échantillonnage aléatoire, comme l'ont fait Cossette (1994) et Tweedie et Baayen (1998). Leur procédure est employée dans la présente recherche à deux différences près. Tout d'abord, au lieu d'être appliquée à un ou deux textes, elle le sera à plus de 600 textes rédigés par des apprenants d'une deuxième langue. Analyser un grand nombre de textes permet de vérifier que l'effet de la longueur sur un indice est similaire pour de nombreux textes, ce que les analyses infirmeront. En second lieu, les longueurs d'échantillons comparées sont nettement plus courtes afin d'apporter des informations pertinentes pour l'évaluation de la qualité de textes puisque, dans ce domaine, les textes font rarement plus de quelques centaines de mots. De plus, les analyses porteront également sur des indices développés après la réalisation de ces deux études.

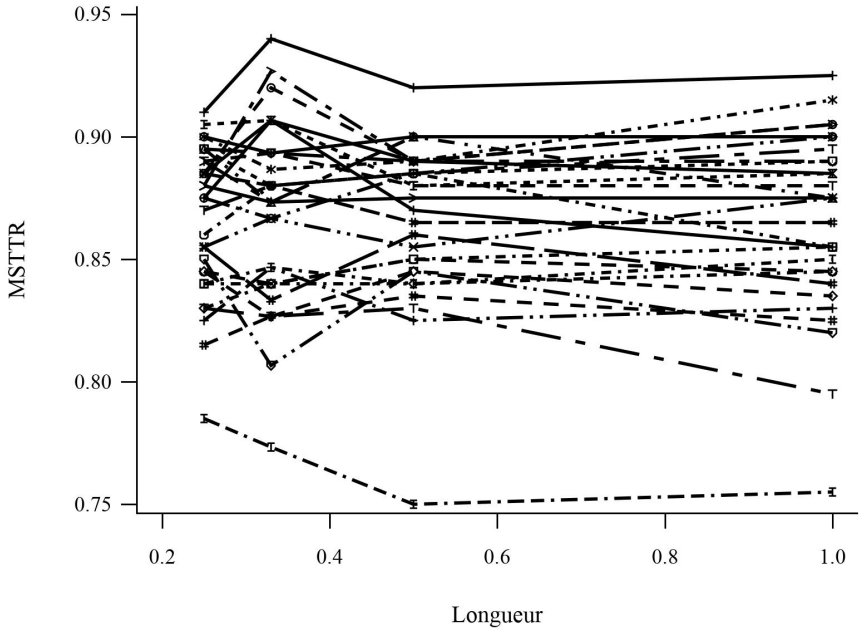

Figure 1. Profils individuels de 30 textes obtenus par la méthode d'échantillonnage parallèle.

Le deuxième problème auquel font face les études qui ont employés la méthode d'échantillonnage parallèle réside dans la procédure d'analyse des résultats. Employer un test d'hypothèse dans le but d'accepter l'hypothèse nulle ne peut pas être recommandé parce que cela fait l'impasse sur la question de la puissance du test et sur les erreurs de type II. De plus, ce test n'évalue que la présence d'un biais systématique comme l'illustre l'analyse suivante, réalisée sur la base d'une partie du matériel qui sera analysé dans les sections suivantes. La méthode d'échantillonnage parallèle a été appliquée aux scores MSTTR des 30 textes en tchèque qui comprennent au moins 240 mots en les tronquant à cette longueur et en les divisant en deux segments de 120 mots, trois de 80 mots et quatre de 60 mots. L'analyse de variance pour mesures répétées indique que la longueur des segments n'a pratiquement pas d'impact sur les scores : $F(3, 87) = 0{,}28$, $p = 0{,}84$. La figure 1 présente les profils individuels des 30 textes. Manifestement, les scores de DL de nombreux textes varient fortement selon la longueur de l'extrait analysé, indiquant une sensibilité de cet indice à la longueur lorsque la méthode

d'échantillonnage parallèle est employée. Le test *F* est non significatif parce que l'effet de la longueur n'est pas le même pour tous les textes. Il conduit à une conclusion erronée.

Il est donc nécessaire de remplacer le test de signification par une mesure de fiabilité qui répond à la question de savoir si un indice attribue toujours les mêmes scores de DL à un texte quelque soit la longueur des extraits comparés. Tel est la fonction du coefficient de corrélation intra-classe (CCI) d'accord absolu (McGraw et Wong, 1996) puisqu'il mesure la part de variance dans un score de DL qui peut être attribuée aux différences entre les textes. Si cette part de variance est proche de 1, on peut en conclure que l'indice n'est pas (ou pratiquement pas) affecté par la longueur des échantillons.

Dans l'étude empirique rapportée ici, l'approche basée sur la méthode d'échantillonnage parallèle sera donc remplacée par la procédure d'échantillonnage aléatoire et le CCI. Avant de présenter cette étude, il est à noter que, si presque tous les chercheurs dans ce domaine ont employé l'approche décrite ci-dessus, les travaux de Fergadiotis et collaborateurs (Fergadiotis, Wright et Green, 2015; Fergadiotis, Wright et West, 2013) s'en distinguent et sont fréquemment cités pour rejeter HD-D et donc l'indice de Muller. Ces auteurs ont utilisé une approche indirecte basée sur des techniques d'analyse multidimensionnelle (analyse factorielle et modélisation par équations structurelles). Ces techniques visent à déterminer dans quelle mesure différents indices sont sous-tendus par la DL, considérée donc comme une variable latente. Il est bien établi que les résultats de ce genre d'analyses sont influencés par les variables qui sont sélectionnées. Si certains indices sélectionnés sont sensibles à la longueur des textes, ils introduiront une part de variance contaminée par ce facteur. Si certains indices sont plus ou moins fortement corrélés entre eux, ils tireront la solution vers ce qu'ils ont en commun. Or, Fergadiotis, Wright et Green (2015) ont analysé deux indices très proches (MATTR et MTDL), l'indice de Maas connu pour sa sensibilité à la longueur des textes, même s'il l'est moins que le TTR, et la version de Malvern et al. (2004) de HD-D. Analyser l'impact de ces choix sur les conclusions de l'étude est indispensable avant de pouvoir tirer argument de leurs observations.

3. Étude empirique

Cette étude évalue la sensibilité à la longueur des textes d'une série d'indices classiques en lexicométrie et/ou recommandés dans les recherches en évaluation du langage. Les analyses portent sur trois langues : l'allemand, l'italien et le tchèque.

Comme expliqué dans l'introduction, les indices basés sur des lois de probabilité, comme l'indice de Muller ou HD-D, sont insensibles par construction à la longueur des textes comparés. Les analyses empiriques ne peuvent donc que confirmer cette insensibilité d'autant plus que la procédure d'échantillonnage aléatoire approxime la loi de probabilité sous-jacente à ces indices (Cossette, 1994). Toutefois, le nature aléatoire de l'échantillonnage induit une part d'erreur. HD-D et l'indice de Muller serviront donc de points de repère pour jauger cette erreur non induite par les indices. La suite de cette section décrit les indices évalués, le matériel analysé et la procédure employée avant de présenter les résultats obtenus.

3.1. Indices évalués

Les treize indices suivant ont été évalués. Dans les formules, N représente le nombre de mots, V le nombre de mots différents, V_i le nombre de mots différents qui sont présents i fois dans un texte et k la fréquence du mot le plus fréquent alors que n représente la taille d'un échantillon extrait ou d'un segment de texte (fenêtre mobile).

Neuf indices sont basés directement sur N et/ou V : le TTR (V/N), le R de Guiraud ($V/\sqrt{N}$), le C de Herdan ($\log V/\log N$), le k de Dugast ($\log V/\log\log N$), le s de Summer ($\log\log V/\log\log N$), le U de Uber ($\log N \times \log N/(\log N - \log \mathrm{V})$), le a de Maas ($\sqrt{(\log N - \log V)/(\log N \times \log N)}$), le H de Honoré ($100 \log N/(1 - V_1/V)$) et le S de Sichel ($V_2/V$). L'indice W de Brunet ($N^{V^{\alpha}}$, 1978) n'est pas évalué parce que les premières analyses ont montré que la valeur de son paramètre α avait un impact très important. La valeur habituelle de -0,185 a donné lieu à des performances très faibles alors que des valeurs dans la zone -0,250 à -0,280 ont produit des performances supérieures aux indices présentés ci-dessus.

Deux indices sont basés sur des lois de probabilité : l'indice binomial de Muller ($V - \sum_{i=1}^{k}(1 - n/N)^i V_i$) et HD-D ($\sum_{i=1}^{k} V_i\ (1 - Hyp(0, N, n))$ où *Hyp* est la fonction de masse de la loi hypergéométrique qui donne la probabilité qu'un mot présent V_i fois dans un texte de *N* mots soit absent d'un échantillon de *n* mots). Dans les expériences, le paramètre *n* est fixé à 50.

Les deux derniers indices se distinguent des autres par le fait qu'ils constituent des algorithmes de calcul et non des formules mathématiques. MSTTR est obtenu en divisant un texte en segments contigus de *n* mots et calculant la moyenne des TTR de chacun de ces segments (Malvern et al., 2004). Le paramètre *n* est habituellement fixé à 50. Il est important de noter que si la division de *N* par *n* ne produit pas un nombre entier, le cas le plus fréquent, les *Mod(N,n)*n* derniers mots du texte ne sont pas pris en compte dans le calcul. MATTR est obtenu en faisant glisser une fenêtre de *n* mots tout au long du texte, en commençant avec le premier mot et en avançant chaque fois d'un mot. Le score final correspond à la moyenne des TTR calculés sur chaque fenêtre (Covington et McFall, 2010). Comme pour MSTTR, *n* est habituellement fixé à 50 en évaluation de textes. MATTR étant basé sur une moyenne mobile, les *n-1* premiers et derniers mots interviennent moins que les autres dans le score puisqu'ils sont présents dans un plus petit nombre de fenêtres.

3.2. Méthode

3.2.1. Matériel

Les textes employés dans cette étude sont extraits du corpus MERLIN (Boyd et al., 2014). Ce corpus, librement disponible, contient 2 267 textes, comme des lettres ou des courriels, rédigés par des apprenants de l'allemand (DE), de l'italien (IT) et du tchèque (CZ) langues secondes. Les textes ont été segmentés en mots et catégorisés grammaticalement par les auteurs du corpus au moyen d'une procédure automatique basée sur le système UIMA (uima.apache.org). Les noms propres, nombres et symboles n'ont pas été pris en compte dans les analyses qui suivent. La longueur des textes est très variable, certains comptant plus de 400 mots et d'autres moins de 10. Afin de pouvoir analyser la stabilité des indices pour des longueurs d'extraits variant dans une plage suffisamment grande, seuls les textes contenants au moins 150 mots ont été conservés soit 182 CZ, 292 DE et 154 IT.

3.2.2. Procédure d'échantillonnage aléatoire

Des échantillons de 60 à 140 mots par pas de 20 ont été extraits de chaque texte au moyen d'une procédure de sélection aléatoire. Cinq mille échantillons indépendants ont été extraits de chaque texte pour chaque longueur ; le score final d'un indice pour un texte et une longueur est la moyenne des scores de ces 5 000 échantillons. Comme expliqué, les recherches visant à évaluer des indices de DL au moyen de la méthode d'échantillonnage parallèle tronquent tous les textes analysés à une même longueur de manière à garantir que les segments obtenus après division contiennent toujours le même nombre de mots. Lorsque la procédure de sélection aléatoire est

employée, il n'y a aucune raison de procéder de cette manière, la procédure ramenant tous les textes à une même longueur. Ne pas tronquer les textes semble d'ailleurs préférable puisque cela correspond à la situation habituelle dans laquelle les textes comparés sont de longueurs différentes. Dans la présente étude, les textes ont donc été analysés dans leur intégralité.

3.3. Analyses et résultats

Les figures 2, 3 et 4 présentent les CCI pour les trois langues. Le CCI moyen est représenté par un cercle et les barres indiquent un intervalle de confiance à 95% pour la valeur du CCI dans la population (McGraw et Wong, 1996). Les indices sont ordonnés du plus petit au plus grand CCI pour l'italien. Ces figures confirment, si cela était nécessaire, que tant l'indice de Muller que HD-D ne sont pas affectés par la longueur des textes comparés. Elles montrent aussi que la variabilité liée à la procédure de sélection aléatoire des mots est extrêmement faible, le CCI moyen pour l'indice de Muller étant supérieur à 0,9996 dans les trois langues.

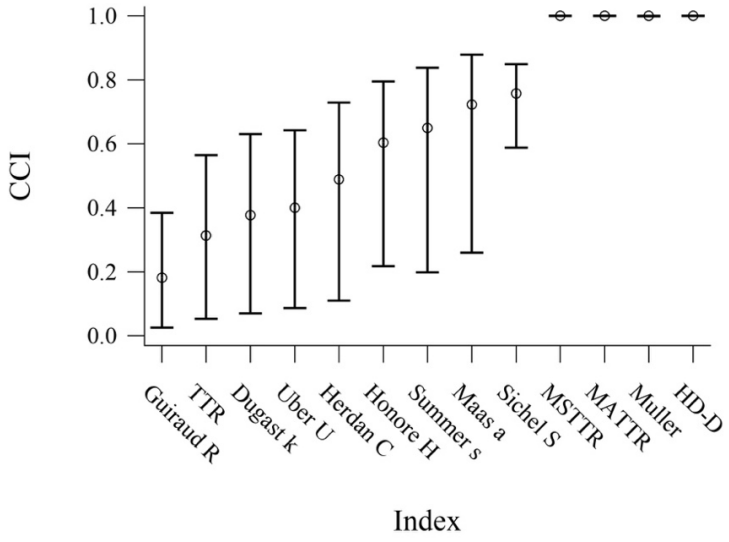

Figure 2. CCI moyens et intervalles de confiance pour l'italien.

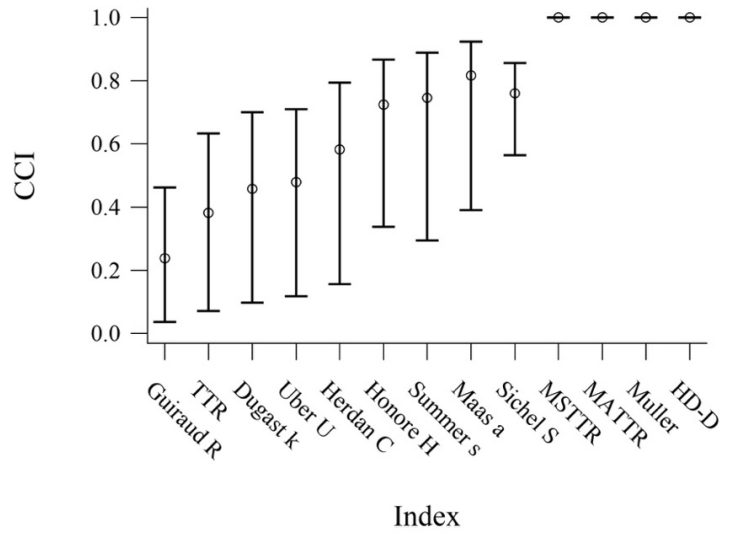

Figure 3. CCI moyens et intervalles de confiance pour l'allemand.

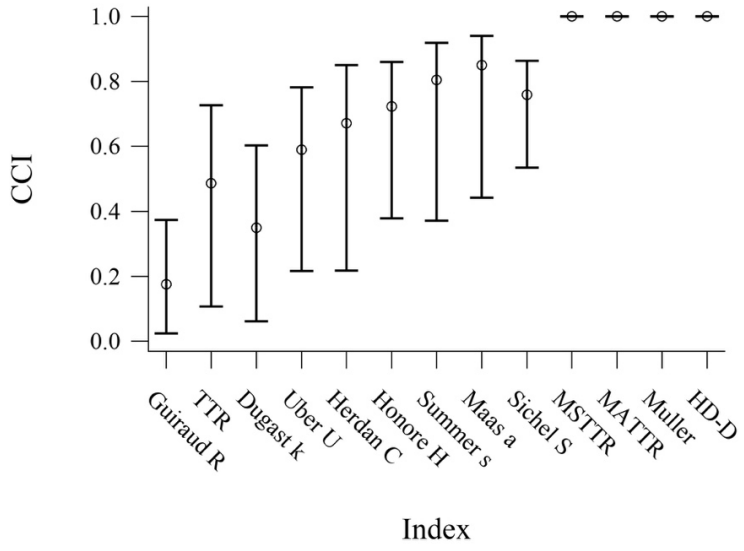

Figure 4. CCI moyens et intervalles de confiance pour le tchèque.

Ces figures indiquent également que MATTR et MSTTR sont insensibles aux différences de longueur des textes. Ce n'est pas le cas des autres indices, les meilleurs d'entre eux atteignant à peine un CCI de 0,80. De plus, tous les intervalles de confiance pour ces indices incluent des valeurs très faibles. Le S de Sichel obtient les résultats les moins médiocres ainsi qu'un intervalle de confiance beaucoup plus petit que ceux des autres indices sensibles à la longueur. Néanmoins, ces résultats indiquent qu'aucun de ces indices ne devrait être employé en évaluation du langage.

Afin de rendre plus concret l'impact de la longueur sur les indices, la figure 5 présente les profils de chaque texte en italien pour six indices. Des graphiques très similaires ont été obtenus avec les textes des deux autres langues. Afin de rendre cette figure aussi informative que possible malgré le grand nombre de profils représentés, les douze profils, qui comptent le plus de différences très positives ou très négatives entre des longueurs sont rendus plus visibles. Le même critère et le même algorithme sont appliqués à tous les indices de manière à ne pas favoriser certains d'entre eux.

Les profils pour l'indice de Muller, et donc aussi HD-D, sont presque parfaitement stables; il en est de même pour MSTTR (et MATTR qui n'est pas représenté faute de place). Par contre, le R de Guiraud présente clairement un biais croissant lié à la longueur alors que le TTR présente un biais décroissant. Les profils obtenus avec le H de Honoré et surtout le S de Sichel montrent des croisements importants entre les courbes, contrairement aux autres indices. Cette observation, inattendue pour moi, semble pouvoir être mise en relation avec le fait que ces deux dernières mesures sont les seules parmi les treize évaluées qui utilisent la fréquence des hapax legomena et des dis legomena en plus du nombre total de mots différents.

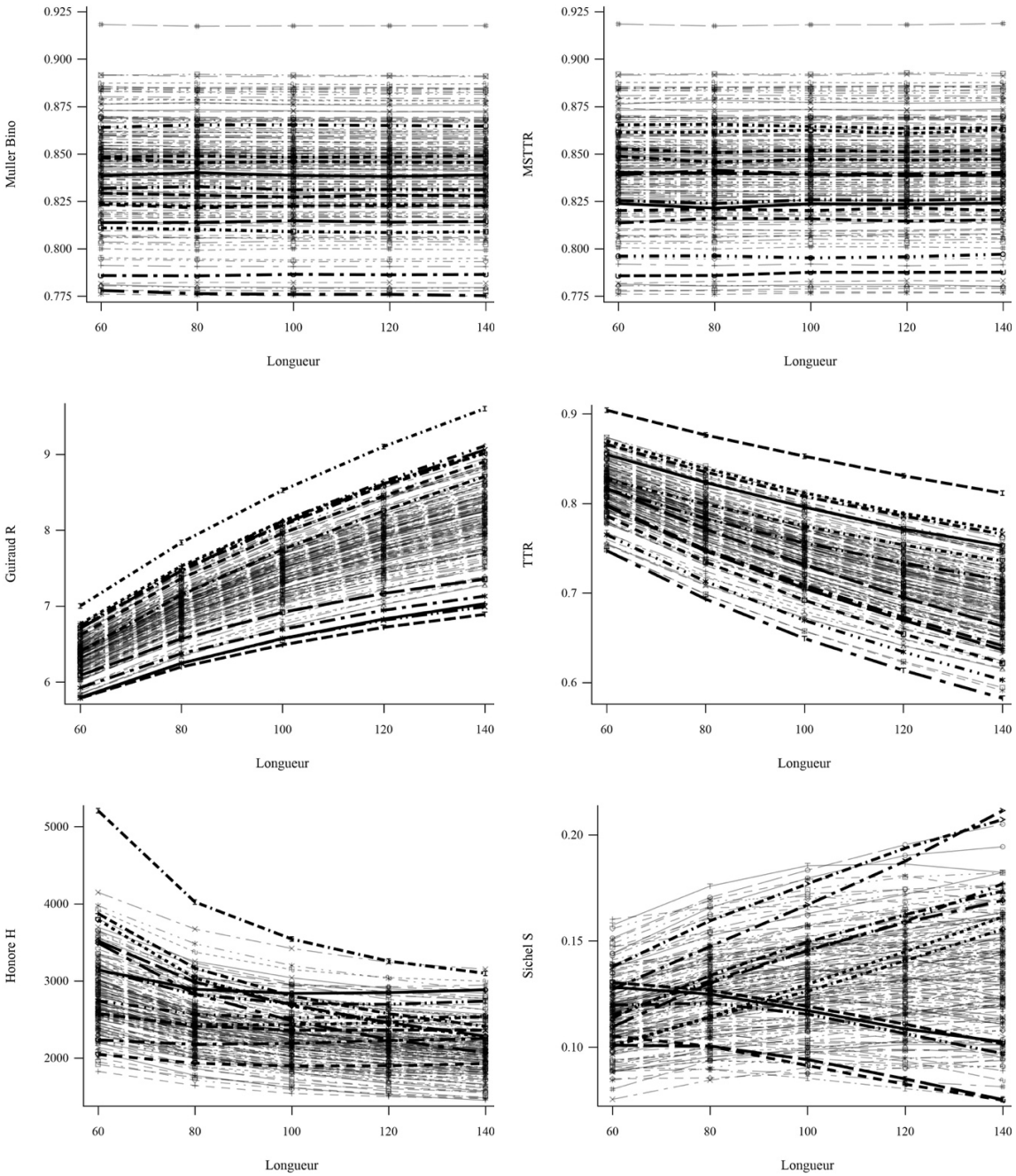

Figure 5. Profils individuels pour six indices sur la base des textes en italien.

4. Discussion et conclusion

Cette recherche porte sur la sensibilité d'indices de DL à la longueur des textes. Son objectif est de déterminer quels indices parmi les treize évalués peut être employé pour comparer des textes de longueurs différentes. Il est important de souligner que l'objet est la DL d'un texte complet et non l'évolution de celle-ci tout au long d'un texte ou la comparaison de la DL d'un texte avec celle de sections de celui-ci, des questions qui ont aussi reçu l'attention en lexicométrie. Les analyses effectuées confirment et complètent les observations de Cossette (1994). Les indices

basés sur une loi de probabilité sont parfaitement stables. Ces analyses montrent aussi que MATTR et MSTTR sont également insensibles aux différences de longueur lorsqu'on les analyse au moyen de la procédure utilisée ici. Ces deux indices ne reposent pourtant par sur une loi de probabilité. Ils présentent cependant une propriété commune avec l'indice de Muller : ils estiment la DL sur la base de segments de la même taille dans tous les textes analysés. En d'autres mots, au lieu de ramener tous les textes à une même longueur par une approche probabiliste, ils le font concrètement. MATTR et MSTTR se distinguent des deux indices probabilistes par deux propriétés. La première est qu'ils estiment la DL sur la base du nombre de mots différents dans des segments de mots contigus et non sur la base de la distribution des mots différents dans l'ensemble du texte. Ils permettent donc une analyse de l'évolution de la DL dans un texte. D'un autre côté, comme expliqué avant, ils ne prennent pas en compte de la même manière tous les mots d'un texte. Lorsque l'évolution dans le texte n'est pas pertinente, il semble donc préférable d'employer les indices basés sur une loi de probabilité.

Les analyses mettent aussi en évidence des différences de sensibilité à la longueur entre les autres indices, mais aucun de ceux-ci ne semble suffisamment stable pour pouvoir concurrencer les quatre meilleurs. Il est toutefois important de garder à l'esprit que tous les résultats présentés ont été obtenus avec des textes courts puisque composés de 150 à 400 mots. Il s'agit là de longueurs courantes en évaluation du langage (Bestgen, 2017, 2020), mais il est possible que certains de ces indices soient plus performants lors de l'analyse de textes beaucoup plus longs.

Analyser des textes plus longs est certainement une piste pour de futures recherches. Il serait aussi intéressant d'essayer de comprendre l'absence de parallélisme des profils pour le H de Honoré et le S de Sichel. Les résultats préliminaires pour le W de Brunet mériteraient aussi une analyse. Par contre, des résultats très similaires ayant été obtenus (comme attendu) pour les trois langues, il ne semble pas prioritaire de se lancer dans l'analyse de langues supplémentaires.

Remerciements

L'auteur est chercheur qualifié du F.R.S-FNRS.

Bibliographie

Baayen R.H. (2001). Word frequency distributions. Springer.

Bestgen Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, 69, 65-78.

Bestgen Y. (2020). Reproducing monolingual, multilingual and cross-lingual CEFR predictions. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, Marseille, 5597-5604.

Bestgen Y. (2024), Measuring Lexical Diversity in Texts: The Twofold Length Problem. *Language Learning*. https://doi.org/10.1111/lang.12630

Boyd A., Hana J., Nicolas L., Meurers D., Wisniewski K., Abel A., Schone K., Stindlova B. et Vettori C. (2014). The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, 1281-1288.

Brunet E. (1978). *Le vocabulaire de Jean Giraudoux. Structure et évolution*. Genève-Paris : Slatkine-Champion.

Cossette A. (1994). *La Richesse lexicale et sa mesure*. Genève-Paris : Slatkine-Champion.

Covington M.A. et McFall J.D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17 (2), 94-100.

Fergadiotis G., Wright H.H. et Green S.B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, 58 (3), 840-852.

Fergadiotis G., Wright H.H. et West T.M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, 22 (2), S397-S408.

Hess C.W., Haug H.T. et Landry R.G. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research*, 32 (3), 536-540.

Hubert P. et Labbé D. (1988a). Note sur l'approximation de la loi hypergéométrique par la formule de Muller. In Labbé D., Serant D. et Thoiron P. (Eds.), *Études sur la richesse et la structure lexicales*. Genève-Paris : Slatkine-Champion, 77-91.

Jarvis S. et Daller M. (2013). *Vocabulary knowledge: Human Ratings and Automated Measures*. Benjamins.

Koizumi R. et In'nami Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40 (4), 554-564.

Malvern D., Richards B., Chipere N. et Durán P. (2004). *Lexical Diversity and Language Development : Quantification and Assessment*. Palgrave MacMillan.

McCarthy P. M. et Jarvis S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24 (4), 459-488.

McGraw K. O. et Wong S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1 (1), 30-46.

Muller C. (1964a). Calcul des probabilités et calcul d'un vocabulaire. Reproduit *dans Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 167-176.

Muller C. (1964b). Étude de statistique lexicale : *L'illusion comique* de Pierre Corneille. Paris : Klincksieck.

Nasseri M. et Thompson P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, 47, Article 100511.

Serant D. (1988). À propos des modèles de raccourcissement de textes. In Labbé D., Thoiron P. et Serant D. (Eds.), *Études sur la richesse et la structure lexicale*. Genève-Paris : Slatkine-Champion, 115-124.

Thoiron P. (1986). Indice de diversité et mesure de la richesse lexicale. In *Méthodes quantitatives et informatiques dans l'étude des textes*. Genève-Paris : Slatkine-Champion, 832-839.

Treffers-Daller J. (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. In Jarvis S. et Daller M. (Eds.), *Vocabulary knowledge: Human Ratings and Automated Measures*. Benjamins, 79-104.

Treffers-Daller J., Parslow P. et Williams S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, 39 (3), 302-327.

Tweedie F. J. et Baayen R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers & the Humanities*, 32, 323-352.

Xanthos A. (2013). L'évaluation (de l'évaluation) de la diversité lexicale. *Cahiers de l'ILSL*, 36, 231-252.

Zenker F. et Kyle K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, Article 100505.